WHITE PAPER

Delivering Optimized Performance for Apache™ Hadoop®

SGI® InfiniteData™ Cluster Solution Architecture and Benchmarks Using Cloudera® Software

TABLE OF CONTENTS

Introduction

Today's enterprise is awash in data. From Internet traffic, sensors, and credit card activity, to social media, video monitors, and more, data is arriving in volume, with velocity, and of variety far greater than ever experienced. What has forward-looking businesses and government agencies focused on Big Data is the value lying within that data, derived through analytics. Online retailers can better understand buying activity to create more effective offers. Hospitals can tailor cancer treatment to patient genetic profiles to increase success. Auto makers can learn more about driving behavior to build safer cars. From competitive advantage and top-line growth to saving lives, the potential gains surrounding Big Data are far reaching but also present new challenges: how to derive value at greater speed, scale, and efficiency.

High-performance computing (HPC) has historically been utilized for complex and computationally intensive problems ranging from scientific discovery, to physical simulations, to government security. To extract value from increasingly massive volumes of digital information, HPC is now being applied to the rapidly emerging field of Big Data analytics. When relationships within data sets are generally understood, enterprises are turning to the open source computing framework of Apache™ Hadoop® software. SGI has been a driving force in HPC for nearly two decades. Drawing upon this expertise building the world's fastest supercomputers, SGI has been at the forefront of Hadoop deployments, with cluster installations now reaching tens of thousands of nodes.

Part of a comprehensive suite of HPC solutions for Big Data, SGI® InfiniteData™ Cluster is designed to meet the growing challenge of Big Data analytics while enabling businesses and government agencies to leverage Hadoop with faster and greater insights at lower cost.

- **Faster performance**. Terasort benchmarks run by SGI consistently demonstrate that SGI InfiniteData Cluster with Cloudera's Distribution Including Apache Hadoop" (CDH) here instead, as it is referring to that specific product for benchmarking delivers the industry's highest performance per rack.
- **Breakthrough efficiency**. Using a unique SGI node architecture, SGI InfiniteData Cluster delivers up to forty cluster nodes and 1.9 PB in a single rack—over twice the compute-storage of any enterprise-class HPC cluster solution.
- **Peta-scalability**. Solutions can start with a small number of cluster nodes in a single rack, or encompass multiple racks with thousands of nodes and linear Hadoop scalability.
- **Countless savings**. Ready to "power up and go", solutions arrive factory-integrated with Cloudera Hadoop software, Red Hat® Enterprise Linux, and SGI Management Center, saving IT departments hundreds to thousands of man hours in setup time, software integration, and testing.

This paper discusses the requirements for optimizing Hadoop and deploying at scale in the enterprise. Details on the SGI InfiniteData Cluster solution architecture running CDH are provided, along with SGI results from the popular Terasort benchmark, including industry comparisons.

Deploying Hadoop with SGI and Cloudera

Hadoop technology has certainly allowed great advances in the ability to understand and extract value from growing amounts of unstructured data. At the same time, expanding analytics demands and ever growing data volumes require enterprise-class solutions delivering the highest levels of performance and scalability with lower total cost of ownership. Combined with Cloudera, SGI InfiniteData Cluster addresses these key drivers for Hadoop infrastructure.

The Challenges of Deploying Hadoop in a Changing World

Hadoop infrastructure requirements are changing, with new demands in terms of speed, scale, availability, and density. Only a carefully designed solution can efficiently provide the level of throughput, processing rate, and round the clock access to information that is needed to fulfill increasing enterprise demands. SGI InfiniteData Cluster combined with Cloudera software provides key advantages to meet this requirement.

- **Solving the challege of Hadoop deployment**. While organizations recognize the value of Hadoop for business analytics, they often struggle to implement Hadoop infrastructure at scale. SGI InfiniteData Cluster enables organizations to rapidly deploy an enterprise-class Hadoop analytics platform as all hardware and necessary software, including operating system, Cloudera software, and SGI Management Center, arrive fully factory integrated. And the flexible architecture enables organizations to scale seamlessly with linear performance to support expanding business applications.
- **Optimizing a Hadoop cluster**. More than merely scaling to larger numbers of cluster nodes, SGI understands that optimizing the Hadoop cluster itself for Big Data workloads is essential. SGI InfiniteData Cluster is pre-built for affordable price/performance, minimizing required floor space, power, cooling, and maintenance for Hadoop infrastructure. This approach combines high-density servers and high-capacity disk drives in an ideal ratio balanced with the number of cores per-node, resulting in linear TeraSort benchmark performance and optimal performance per watt. The solution also includes tested support for high availability, allowing organizations to focus on application development instead of performance tuning and configuration hassles.
- **Supporting increasing volume, velocity, and variety of data**. Organizations need to be able to grow their Hadoop clusters, but they also need effective ways to integrate their data with other systems and move it in and out of the cluster. SGI InfiniteData Cluster has the raw capacity to support hundreds of terabytes to petabytes of high velocity data in high-density rack configurations.

A Strongly Differentiated Approach from SGI

Building on two decades of High Performance Computing (HPC) leadership and experience designing high volume Hadoop clusters, SGI has a strongly differentiated approach to delivering Hadoop infrastructure solutions for the enterprise.

Optimized Hadoop performance and density

Accomplishing better total cost of ownership through infrastructure requires a thorough understanding of the dynamics of Hadoop performance. Organizations need to scale their Hadoop capabilities, but they need to work within the constraints of their data centers in terms of floor space, power, and cooling. To address these concerns, SGI has done the work to understand the ideal balance between compute and storage for Hadoop, and has designed a system that delivers break-though density, both in terms of nodes and capacity. In fact, competitors require twice the footprint or more to deliver the same number of nodes and capacity that SGI InfiniteData Cluster can provide.

As an example of the difficulties faced by typical approaches, traditional 1 rack unit (RU) servers typically can't support the spindles needed for good Hadoop performance, and only 20 traditional 2 RU servers can be supported in a 42 RU rack. In contrast, SGI's 2 RU half-width servers effectively double the capacity in a rack to 40 nodes, and provide the ideal disk to processor core ratio for running Hadoop workloads. SGI optimizes performance and density through an innovative architecture that includes:

- 12 drives per compute node
- An ideal 1:1 di sk to processor core ratio
- High density with 40 nodes and up to 1.92 PB per 42 RU rack

Linear TeraSort Scaling and throughput per rack

These advantages are made clearly apparent with TeraSort benchmark testing performed by SGI and detailed later in this document. The benchmarking employed the SGI IDC3212-RP4 tray based on the six-core Intel Xeon processor E5-2630 in SGI InfiniteData Cluster, together with Red Hat Enterprise Linux and CDH 4.3.1. Not only did this configuration produce linear scalability up to a 24-node cluster, but per-rack TeraSort throughput numbers easily dwarf those of competing infrastructure solutions.

Power up and go

SGI Hadoop solutions are factory integrated with pre-racked, pre-configured, and pre-tested nodes that include networking, Linux, and Hadoop software—all configured and ready to run. This approach can save hundreds to thousands of man hours in deployment. While some competitors have pre-packaged offerings, they can occupy twice the physical footprint or more, driving real estate, power, and cooling costs.

Serviceability

SGI InfiniteData Cluster is designed for data center environments where serviceability is a critical component for lowering operational expenses (OpEx). All InfiniteData Cluster components—trays, interconnects, and I/O—are cold-aisle accessible to support high-efficiency, hot-aisle/cold-aisle floor plans. The server trays pull out easily if in need of service, and disk drives can be added or replaced in minutes.

Hadoop solution architecture on SGI InifiniteData Cluster

Building an effective solution architecture for Hadoop depends on marrying innovative software and system design together with integration that results in a compelling solution.

Solution architecture overview

SGI InfiniteData Cluster is based on a careful combination of software and hardware components that result in high density and very high throughput per rack. Components of the solution architecture are tested and verified to work together, and include:

- SGI IDC3212-RP4 tray servers as Data/TaskTracker nodes (Figure 1) based on the Intel Xeon processor E5-2600 V2 family
- Hadoop NameNode, secondary NameNode, JobTracker, and Application Node based on full-depth SGI C1110-RP6 servers with redundant power supplies, and based on the Intel Xeon processor E5-2600 V2 family
- 10 Gigabit Ethernet networking for high-speed and low-latency
- Software stack using the latest release of CDH 4.3 and Clouder Manager (CM) 4.6
- Red Hat Enterprise Linux 6.4
- Built-in support for high availability, eliminating the NameNode single point of failure common in other Apache Hadoop distributions
- Modular packaging to allow for flexible scalability of Hadoop clusters

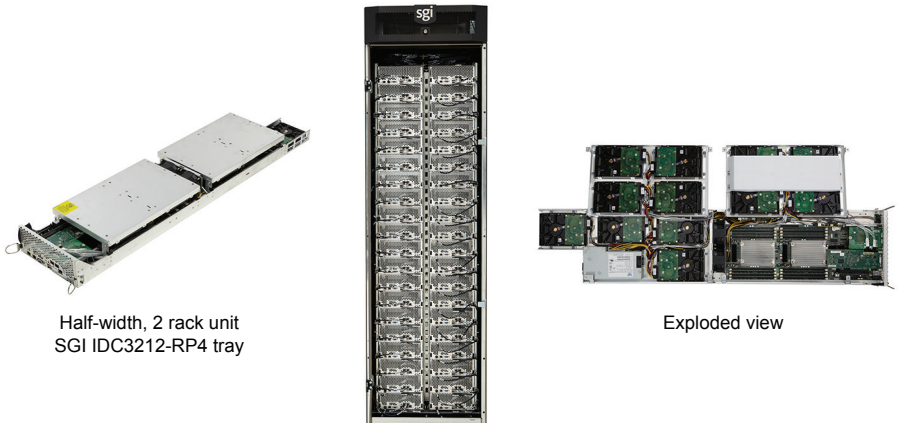

Figure 1. A single rack of SGI IDC3212-RP4 trays can yield 40 Hadoop data nodes and 1.92 PB of capacity.

Hadoop cluster system components

Within SGI InfiniteData Cluster, components are chosen for appropriateness to the task they must perform. DataNodes / TaskTracker Nodes were chosen based on their ability to provide an optimally-balanced configuration of cores, drives, and memory. SGI IDC3212-RP4 half-width server trays offer a 1:1 core to spindle ratio (twelve cores per server with twelve drive spindles), and an affordable cost per terabyte of capacity. The 2 RU server tray contains two half-width servers, each configured with:

- Two Intel® Xeon® Processor EX-2630 V2 (2.6 GHz, six-core)
- Eight 8 GB 1.5v 1866 MHz DIMMs (64GB memory)
- Twelve 3.5-inch 4 TB 7.2K RPM SATA drives in JBOD configuration (48 TB capacity)
- Dual-port 10 Gigabit Ethernet module

NameNode, Standby NameNode, and JobTracker are configured for redundancy and availability for providing key Hadoop services that affect the entire cluster. SGI C1110-RP6 full-depth servers were chosen for density and redundant power supplies, and are configured with:

- Intel® Xeon® Processor ES-2630 V2 (2.6 GHz, six-core)
- Eight 8 GB 1.5v 1866 MHz DIMMs (64 GB memory)
- Four 3.5-inch 4 TB 7200rpm SATA 6 Gb/s drives in RAID 10 configuration
- Dual-port 10 Gigabit Ethernet NIC
- Redundant power supplies

The Application Node is also based on the SGI C1110-RP6 full-depth server. However, the configuration features more processing capability and twice the memory of the NameNodes and JobTracker, and includes:

- Intel® Xeon® Processor ES-2680 V2(2.8 GHz 10-core)
- Sixteen 8GB 1.5v 1866 MHz DIMMs (128 GB memory)
- Four 3.5-inch 4 TB 7200rpm SATA 6 Gb/s drives in RAID 10 configuration
- Dual-port 10 Gigabit Ethernet NIC
- Redundant power supply

Network Interconnection

SGI designed the networking topology for Hadoop clusters with redundancy, scalability, and performance in mind. Each rack has two leaf switches with active-active configuration to provide resilient network connections to the Data/TaskTracker nodes, NameNodes, JobTracker, and Application Node as shown in figure 2. A second Gigabit Ethernet network is provided to each system for an out-of-band management network. To gain utilization, the Standby NameNode is also utilized as the SGI Management Center Administration Node.

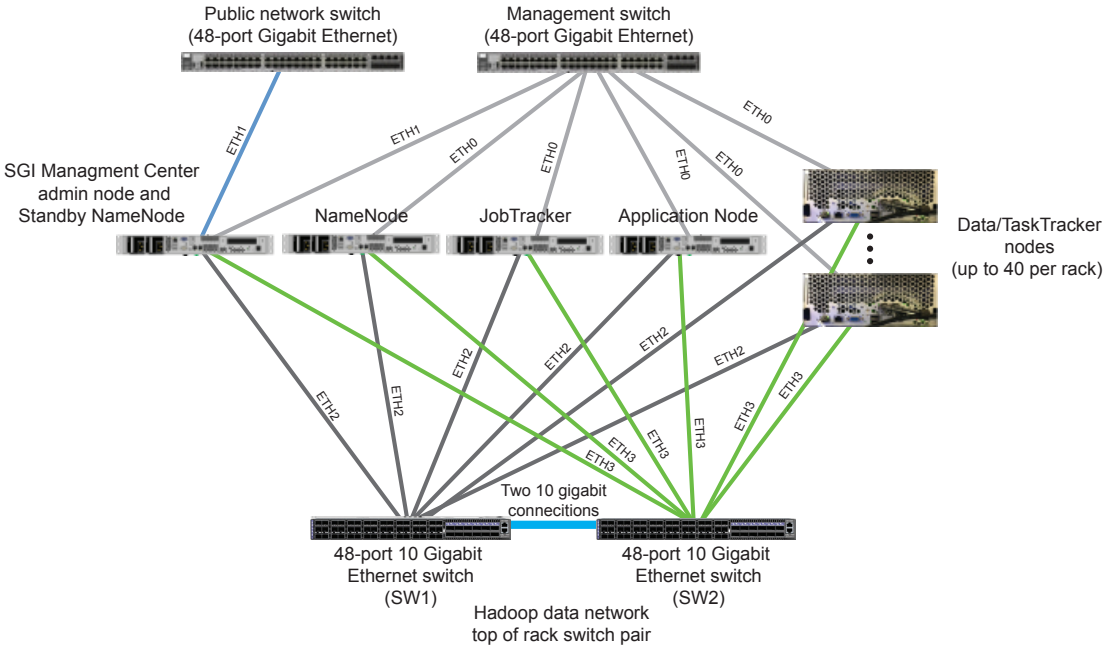

Figure 2. Rack-level network connectivity for SGI InfiniteData Cluster.

Scaling the solution architecture

Organizations need the ability to start small and scale their Hadoop infrastructure as their needs dictate. To serve this need, SGI InfiniteData Cluster employs a flexible and scalable approach that lets Hadoop infrastructure grow incrementally based on the needs of the organization.

A flexible and scalable approach

Hadoop clusters of virtually any size can be built using SGI InfiniteData Cluster. Because this modularity and scalability is designed in from the outset, organizations don't encounter arbitrary limitations and bottlenecks as they grow their infrastructure. SGI InfiniteData Cluster can be built from the following modules:

- An eight-node SGI Hadoop Starter Kit is available for organizations who are new to Hadoop, or who have more modest needs.
- Additional DataNodes can be added incrementally, as required, up to 18 racks and 716 Data/TaskTracker nodes.
- Network infrastructure for the solution architecture is designed to scale along with node count and capacity.

Single rack configuration

As mentioned, SGI IDC3212-RP4 half-width server trays provide considerable density. For example, a single 42 RU rack can support a complete Hadoop cluster with up to 36 Data/TaskTracker nodes. A single-rack configuration includes:

- Four master nodes (NameNode, Standby NameNode, JobTracker, and Application Nodes)
- Up to 36 Data/TaskTracker nodes
- Dual 48-port 10 Gigabit Ethernet switches for the Hadoop data network
- 48-port Gigabit Ethernet management switch for SGI Management Center

MultiRack configuration

By adding additional racks, SGI InfiniteData Cluster can be scaled up to 18 full racks.
Each additional rack adds up to 40 nodes, for a total maximum capacity of up to 716 Data/TaskTracker nodes yielding a capacity of up to over 34 PB of capacity. In multirack configurations, 40 Gigabit Ethernet spine switches connect the dual 10 Gigabit Ethernet switches in each rack.

- Two 36-port 40 Gigabit Ethernet spine switches can support two to nine racks (up to 360 nodes total, including four master nodes, and up to 356 Data/TaskTracker nodes).
- Four 36-port 40 Gigabit Ethernet spine switches can support from 10 to 18 racks (up to 720 nodes total, including four master nodes, and up to 716 Data/TaskTracker nodes).

Figure 3 illustrates multirack connectivity for up to nine racks, showing 40 Gigabit Ethernet spine switches connecting additional racks. Larger cluster configurations (up to 18 racks) require an additional pair of 40 Gigabit Ethernet spline switches.

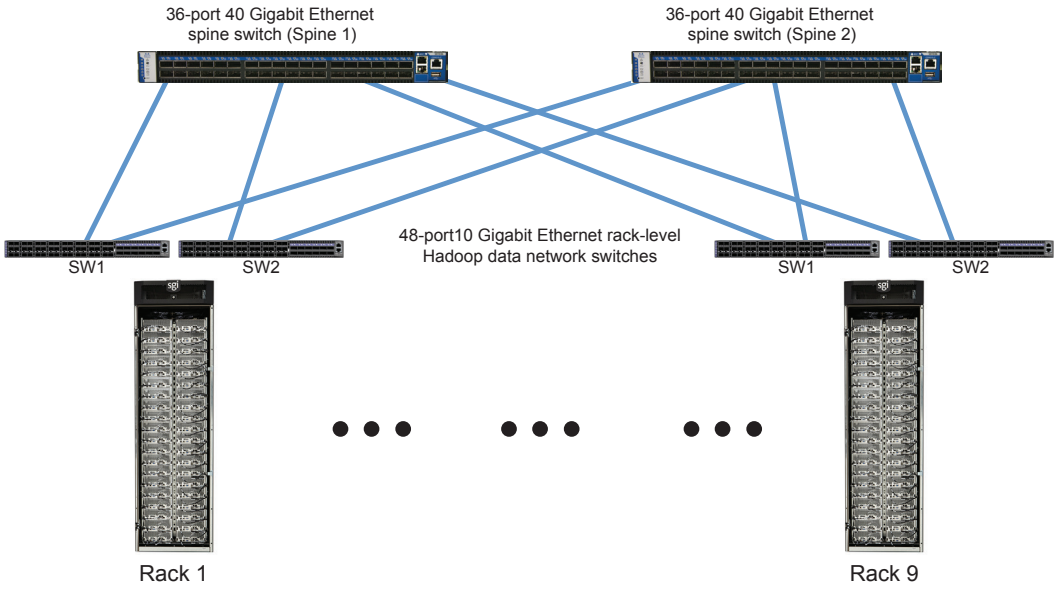

Figure 3. Up to 18 racks can be connected into a single SGI InfiniteData Cluster, yielding up to 716 Data/TaskTracker Nodes, and up to over 34 PB of capacity (nine-rack configuration shown).

Cloudera HDFS HA for high availability

High availability is increasingly critical in Hadoop deployments, as workloads move from traditional Hadoop batch jobs to more interactive and real-time workloads. Unfortunately, the NameNode in many Apache Hadoop distributions represents a single point of failure. If the NameNode system or process became unavailable, the entire cluster can become unavailable until the NameNode is either restarted or brought up on a separate machine. Given the considerable computing resources and large data volumes involved in the typical cluster, this lack of redundancy can dramatically affect utility, reducing the total availability of the cluster in two significant ways:

- In the case of an unplanned event such as a system crash of the NameNode, the cluster is unavailable until an operator restarted the NameNode.
- Planned maintenance events such as software or hardware upgrades on the NameNode system can result in periods of cluster downtime.

Cloudera's HDFS High Availability (HA) feature addresses availability by providing the option of running two redundant NameNodes in the same cluster in an Active/Passive configuration. This functionality allows a fast fail-over to a new NameNode in the case of a system crash. The HDFS HA feature also offers a graceful administrator-initiated fail-over for the purpose of planned maintenance. SGI has performed testing to validate this functionality in the context of SGI InfiniteData Cluster.

Cloudera's quorum-based storage implementation uses Quorum Journal Manager (QJM). In order for the Standby Node to keep its state synchronized with the Active Node in this implementation, both nodes communicate with a group of separate daemons called JournalNodes (Figure 4). When any namespace modification is performed by the Active Node, it durably logs a record of the modification to a majority of these JournalNodes. The Standby Node is

capable of reading the edits from the JournalNodes, and is constantly watching them for changes to the edit log. As the Standby Node sees the edits, it applies them to its own namespace. In the event of a fail-over, the Standby Node will ensure that it has read all of the edits from the JournalNodes before promoting itself to the Active state. This step ensures that the namespace state is fully synchronized before a fail-over occurs.

In order to provide fast failover, it is also necessary that the Standby Node has up-to-date information regarding the location of blocks in the cluster. As such, Cloudera configures the DataNodes with the location of both Acrive and Standby NameNodes. DataNodes, in turn, send block location information and heartbeats to both NameNodes simultaneously.

For the correct operation of a high-availability cluster, it is vital that only one of the NameNodes be active at any given time. Otherwise, the namespace state would quickly diverge between the two, risking data loss or other incorrect results. In order to ensure this property and prevent the so-called "split-brain scenario," the JournalNodes will only ever allow a single NameNode to be a writer at a time. During a fail-over event, the NameNode which is to become active will simply take over the role of writing to the JournalNodes, which will effectively prevent the other NameNode from continuing in the active state, allowing the new Active NameNode to safely proceed with the fail-over process. For more details on the high availability feature in the Cloudera software, please refer to documentation at www.cloudera.com.

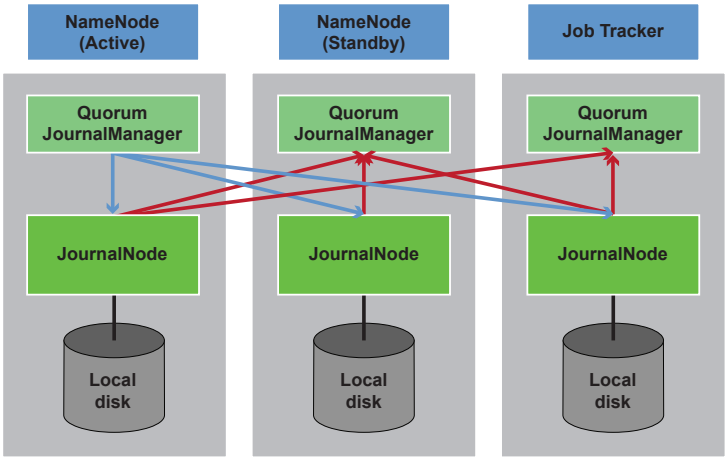

Figure 4. Quorum based high availability in Cloudera software implements a high-availability HDFS.

Leading TeraSort benchmark results

More than merely chasing record results, SGI and Cloudera sought to understand and optimize Hadoop in terms of performance, scalability, and efficiency as it relates to infrastructure.

Performance and throughput

To understand performance of SGI InfiniteData Cluster, SGI conducted various tests around the TeraSort benchmark. While there are no stand benchmarks for evaluating Hadoop performance, TeraSort has emerged as a popular way to exercise the MapReduce functionality Hadoop clusters of various sizes. TeraSort can also be run using a range of data set sizes, and can be a useful metric for comparing the performance and throughput available from a given set of Hadoop infrastructure.

The SGI InfiniteData Cluster has an optimal rack density and throughput per rack, allowing twice the density of competitors or more in a given footprint. As shown in Figure 5, SGI Terasort per-rack throughput for a 10 TB TeraSort dataset[1] is 26% faster than a Cisco UCS® Hadoop cluster[2], and twice as fast as an HP® ProLiant Generation 8 (Gen8) DL380 Hadoop cluster[3], where all three clusters are configured with Intel E5-2600 processors and running Cloudera's distribution including Apache Hadoop.

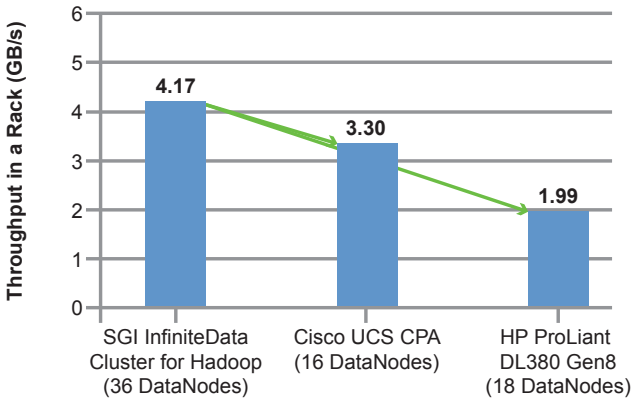

Figure 5. On a per-rack basis, SGI's higher density allows greater throughput than competitors.

Scalability

Scalability is a measure of how appropriate the cluster architecture is for running Hadoop. Cluster infrastructure needs to demonstrate scalability, both as processor resources are increased, and as the cluster node size increases. To evaluate scalability for SGI InfiniteData Cluster, SGI conducted 10 TB TeraSort benchmark tests using different numbers of nodes, as well as different Intel Xeon Processor E5 Series CPUs as shown in Figure 6. Not only did the different numbers of cluster nodes demonstrate strong scalability, but Intel Xeon Processor E5-2600 V2 CPUs demonstrated 10%, 12%, and 15% performance advantages over Intel Xeon Processor E5-2600 CPUs, when running on 8, 16, and 24-node clusters respectively.

1 SGI InfiniteData Cluster (1 rack, 36 DataNodes, each with Intel E5-2630 V2 @2.3 GHz, 64 GB, twelve 1 TB drives)

2 Cisco UCS Common Platform Architecture (1 rack, 16 DataNodes, each with Intel E5-2665 @2.4 GHz, 256GB, twenty-four 1 TB drives)
Source: cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns944/le_tera.pdf

3 HP ProLiant DL380 Gen8 server (1 rack, 18 DataNodes, each with Intel E5-2667 @ 2.9 GHz, 64 GB, sixteen 1 TB drives).
Source: hp.com/hpinfo/newsroom/press_kits/2012/HPDiscover2012/Hadoop_Appliance_Fact_Sheet.pdf

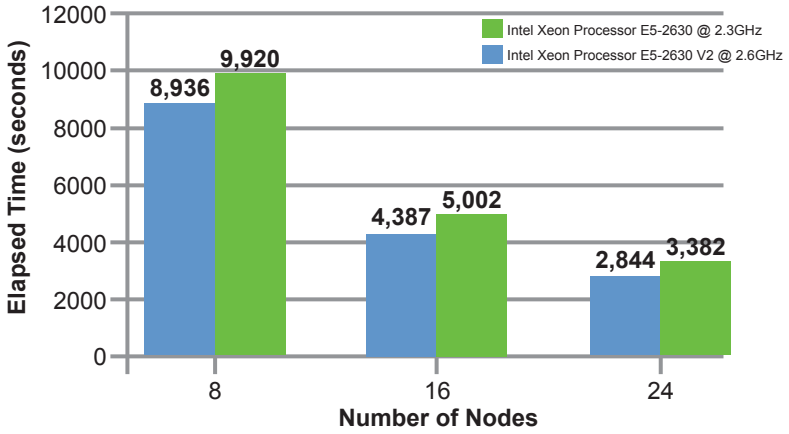

Figure 6. SGI InfiniteData Cluster for Hadoop demonstrates performance scalability in elapsed time (smaller is better) for different sized clusters, as well as performance improvements with faster Intel Xeon Processor E5-2600 V2 CPUs.

SGI testing also measured throughput in terms of megabytes per second during the various 10 TB TeraSort benchmark runs. As with performance testing, throughput was greater for Intel Xeon Processor E5-2600 V2 CPUs, with a 24-node Hadoop cluster generating just under 4 GB/second of throughput (Figure 7).

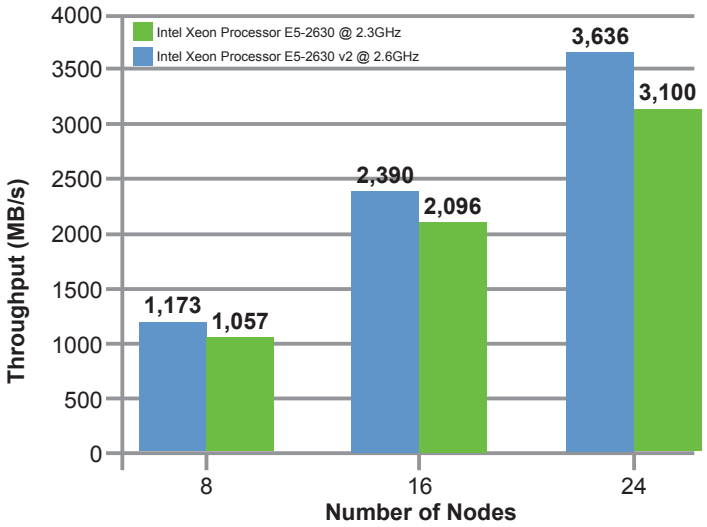

Figure 7. SGI InfiniteData Clsuter for Hadoop exhibits increasing TeraSort throughput with increasing node count as well as improved throughput with faster Intel Xeon Processor E5-2600 V2 CPUs.

In fact, as shown in Figure 8, perfect linear throughput scaling was observed when going from 8, to 16, to 24 nodes within the Hadoop cluster.

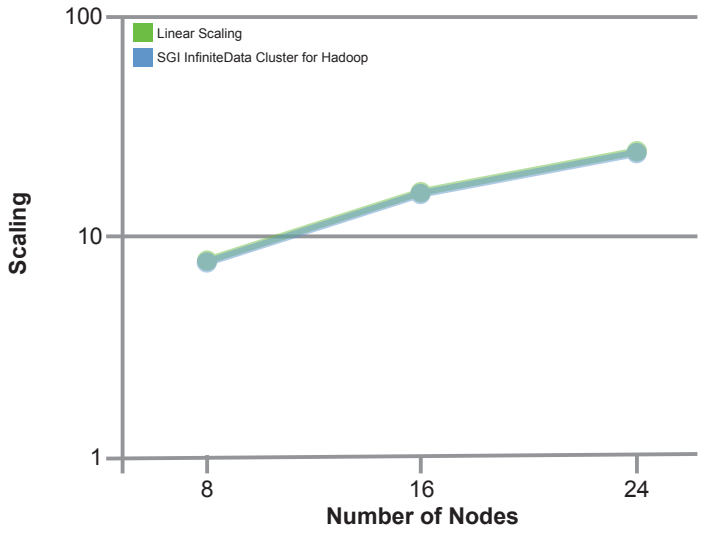

Figure 8. SGI InfiniteData Cluster for Hadoop achieved perfect linear scaling in Hadoop clusters from 8 to 24 nodes.

Efficiency

Organizations deploying Hadoop infrastructure need to be sure that they are getting the most from their infrastructure in terms of performance and throughput. Beyond scaling processor technology and numbers of nodes, organizations want to be sure that they are getting the most performance efficiency from individual systems and the cluster as a whole. For Hadoop workloads, the processor core to drive spindle ratio can make a critical difference in performance efficiency. To evaluate the role of processor core to spindle ratio, SGI ran 10 TB TeraSort benchmarks across a 32-node SGI InfiniteData for Hadoop cluster, varying the number of disk drives for each run.

As shown in Figure 9, results clearly demonstrated that a 1:1 core to drive spindle ratio is important for Hadoop performance. In fact, SGI esting showed a 48% better performance result using SGI InfiniteData Cluster optimized for a 1:1 core to drive spindle ratio.

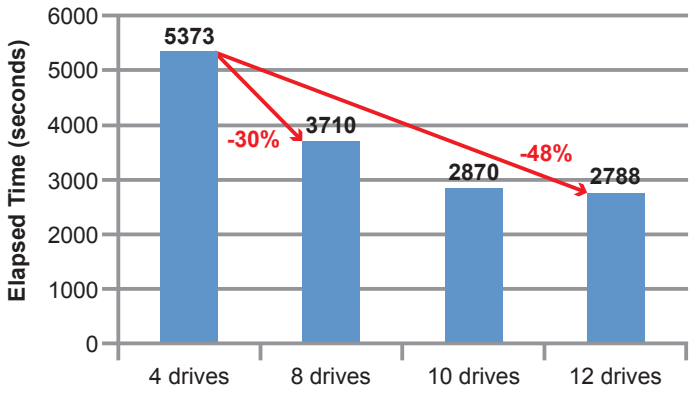

Figure 9. TeraSort @10TB on a 36-node SGI InfiniteData Cluster showing 4, 8, 10, and 12 drives (lower is better), confirming that a 1:1 processor core to disk spindle ratio is ideal for Hadoop workloads.

Conclusion

The increasing value of Big Data analytics in today's enterprise is driving Hadoop infrastructure requirements to new levels. SGI InfiniteData Cluster enables organizations to deploy Hadoop at petascale with predictable, industry leading performance and space efficiency. Arriving pre-racked, pre-configured, and pre-tested with Cloudera Hadoop software running on Red Hat Enterprise Linux provides businesses with a world-class analytics platform in days rather than months, saving IT departments countless hours. And the solution's flexible, highly scalable architecture enables the enterprise to expand Hadoop environments seamlessly for growing data volumes and rising analytic demands.

References

For more information on SGI InfiniteData Cluster, please visit: www.sgi.com/solutions/bigdata/hadoop.

Global Sales and Support: sgi.com/global